다수의 SNS를 이용한 키워드 트렌드 분석 시스템

이명철, 한수현, 이재성
충북대학교 소프트웨어학과
e-mail : lmc1992@naver.com

A Keyword Trend Analysis System Using Multiple SNS Sites

Myung-Chul Lee, Soo-Hyun Han, Jae Sung Lee
Dept of Computer Science, Chungbuk National University

요 약

기업이나 정부 등의 정책 결정에 활용하기 위해, SNS에서 사용하는 키워드를 추출하여 소비자나 유권자의 관심과 선호도를 분석하는 방법이 많이 사용되고 있다. 본 논문에서는 다수의 SNS 사이트에 올린 글과 그에 대한 공감(좋아요) 댓글, 해시태그를 분석하여 관심 키워드의 트렌드를 분석할 수 있는 시스템을 제안한다. 이 시스템에서는 각각의 SNS 글을 형태소 분석하여 키워드 빈도를 측정하고 그에 대한 공감 및 해시태그의 갯수를 계산하여 일정기간 동안의 변화를 그래프로 표시하였다. 이를 통해, 여러 사이트에서의 키워드 트렌드를 한눈에 확인할 수 있도록 했다.

1. 서론

국내 정보통신정책연구원이 발표한 "SNS이용추이분석"[1]에 따르면 조사에 참여한 전체 응답자 10,464명 중 48.2%가 SNS를 사용하는 것으로 나타났다. 이는 2011년 16.8%에서 31.2%가 증가한 수치이다. 또한 앱, 리테일 분석 서비스 와이즈앱[2]에 따르면 2019년 국내 안드로이드 스마트폰 이용자의 앱별 사용 시간을 조사한 결과 네이버, 페이스북 등의 SNS 이용 시간이 2018년보다 38% 증가했다고 밝혔다. 이렇게 연령대를 초월하는 많은 현대인은 소셜 네트워크 서비스인 SNS를 사용하고 있으며, 기업들이 이를 이용하여 사용자의 선호도나 취향을 분석하고 추천 시스템이 다양하게 개발하는데 집중하는 것은 당연하다.

본 논문에서는 위와 같이 기하급수적으로 증가하는 SNS 사용자들의 취향 및 선호도를 분석하기 위해, 소셜 네트워크 서비스에서 크롤링한 사용자의 글과 좋아요 댓글 들을 사용한다. 사용자가 입력하는 검색어와 글에 있는 키워드를 비교하여 빈도를 측정하고, 해시태그와 공감(좋아요)수를 측정하여 이를 그래프로 그려 트렌드를 한눈에 확인할 수 있도록 한다. 본 논문의 2장에서는 관련 연구에 대해서 살펴본다. 3장에서는 트렌드 분석을 위한 키워드를 추출하는 방법과, 해시태그 및 공감(좋아요) 수를 측정하는 방법을 기술하고, 4장에서는 시스템의 구현 및 결과에 관해 기술한다. 마지막으로 5장에서는 결론 및 향후 연구 방향에 대해서 기술한다.

2. 관련연구

인터넷의 보급 및 발달로 온라인 커뮤니티와 SNS를 통해 인간관계를 유지해 나가고 있다. 이러한 SNS의 시장 점유율의 급격한 발달로 인하여 시중에는 SNS 빅데이터를 이용한 소비자 성향을 분석하는 방법, 트위터 데이터를 활용한 IT 트렌드 분석 등 많은 시스템이 개발되어 있다. 또한 계속해서 SNS의 특성을 이용해 트렌드를 분석하는 방법이 연구되고 있다.

기존 시스템들은 사용자의 검색기록을 통해 수치를 비교하고, 이를 통계치를 내어 트렌드를 분석하고 있다[1]. 이러한 시스템들은 시스템을 사용하는 사용자들을 기반으로 현재의 트렌드를 분석할 수 있다는 장점이 있다. 하지만 하나의 사이트만을 이용하는 사용자의 검색어에 대한 정보만이 분석 가능하여 다소 편향된 분석을 보일 수도 있는 단점이 있다. IT 트렌드와 트위터 데이터를 수집하여 도출된 트렌드 분석을 비교한 결과 강한 양의 상관관계를 가지는 것으로 도출 된 기존 연구는 표본 내의 검증만 이루어졌다[3]. 이는 트위터 데이터의 수집이 한정적이므로 정확도의 여부가 불투명하다.

본 논문에서는 사용자가 한눈에 트렌드 추이를 확인할 수 있도록 다수의 SNS 계정을 이용하여 트렌드 분석을 진행하였다. 또한 보다 많은 정보를 포함시켜 사용자에게

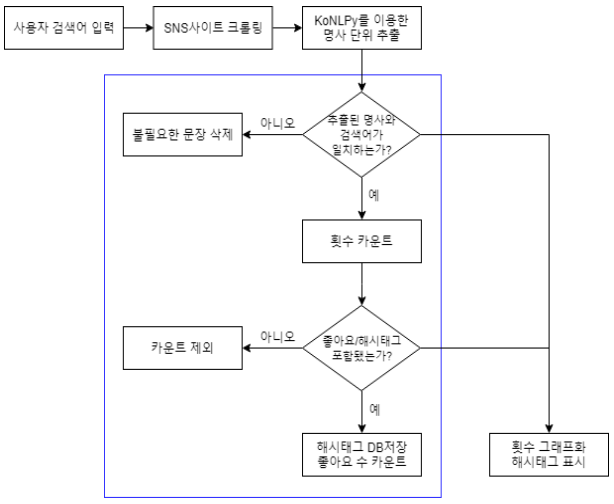
1) http://www.kisdi.re.kr/selectSingleBoard/1319899
2) https://www.wiseapp.co.kr/dashboard/

편리함을 제공하기 위해 SNS의 해시태그 및 공감(좋아요), 공유를 이용하는 방법들을 제안한다.

3. 트렌드 분석을 위한 명사단위 정보 추출

3.1 시스템 구조

(그림 1)은 시스템의 전체적인 동작 절차를 나타내는 도면으로 도면에 표시된 각각의 단계를 거치며 프로그램이 진행될 수 있도록 한다. 사용자가 입력하는 검색어를 기반으로 분석을 진행할 사이트의 일정기간의 글들을 크롤링해오도록 한다. 이후 크롤링된 글을 파이썬 NLP(Natural Language Processing)인 KoNLPy를 사용해 명사 단위로 추출한다. 추출된 명사들은 사용자가 입력한 검색어와 대조하여 일치하는 글의 횟수만을 카운트한다. 일치하지 않는 문장들은 불필요한 문장으로 분류하여 삭제한다. 일치하는 키워드가 존재하는 글만을 가지고 와 글에 공감(좋아요)이 포함되어 있는지를 확인한다. 만약 글에 공감 및 좋아요가 포함된 경우에는 버튼이 눌린 횟수를 카운트하여 데이터베이스에 저장한다. 이 결과 값은 날짜를 기반으로 그래프로 그려 사용자에게 제공한다.

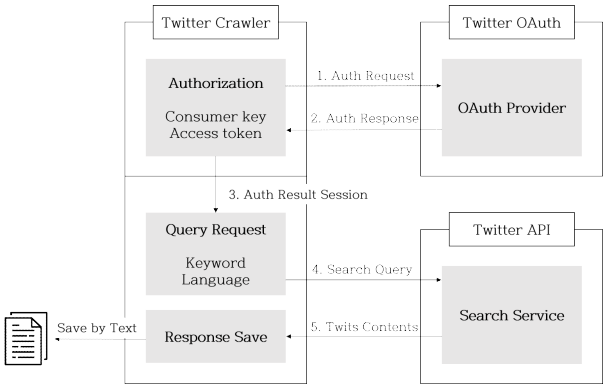


(그림 1) 시스템 흐름도

3.2 프로그램 수행 구조

사용자의 검색어가 포함된 트윗들을 수집하기 위해서 트위터에서 제공하는 Rest 방식의 API를 활용한다. 본 시스템은 API를 통해 키워드 쿼리를 전송하고, 응답으로 전송되는 트윗을 저장하는 크롤러 역할을 포함하고 있다. 해당 API를 통해 사용자가 설정한 쿼리에 대해 질의하고 응답하는 결과는 텍스트 파일로 저장하고 있으며 저장하는 문서의 템플릿을 제공하는 형태로 구성하였다.

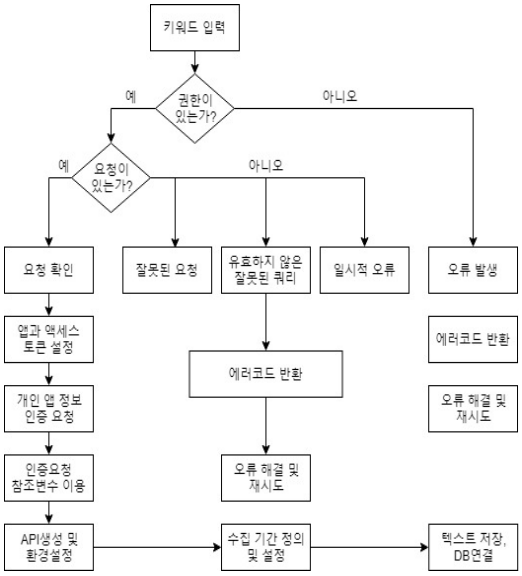
트위터의 API를 이용하기 위해서는 아이디, 비밀번호가 아닌 컨슈머 키와 액세스 토큰을 이용해야 한다. 해당 키와 토큰은 트위터의 API URL[3]을 통해 Application을 등록하여 발급받는다. 페이스북은 Page API를 통해 액세스 토큰을 받는다. 그리고 post id를 가져와 post message로 게시글의 데이터를 가져오고 post reactions를 통해 좋아요 수를 체크한다. 인스타그램은 별도의 API가 없어 URL 페이지로 들어가 크롤링한다.



(그림 2) 트위터 API를 활용한 프로그램 수행 구조

3.3 텍스트 전처리 및 그래프화

SNS API를 사용하여 가져온 텍스트 파일은 KoNLPy를 활용하여 형태소 분석을 진행한다. KoNLPy로 분석된 형태소 중에서 명사 단위의 키워드만을 추출하여 횟수를 날짜별로 카운트하고 그 카운트 횟수를 몽고DB 안에 저장한다. DB에 저장된 데이터는 각각의 SNS별로 서로 다른 색상으로 그래프화 되어 사용자에게 보인다. 또한 해시태그와 좋아요, 공감(좋아요)수와 단어의 노출빈도를 이용하여 그래프를 보여주어 단어 출력의 빈도의 정확도를 높여준다. 그래프의 형태는 꺾은선 그래프로 나타낸다.



(그림 3) 크롤링 프로그램 순서도

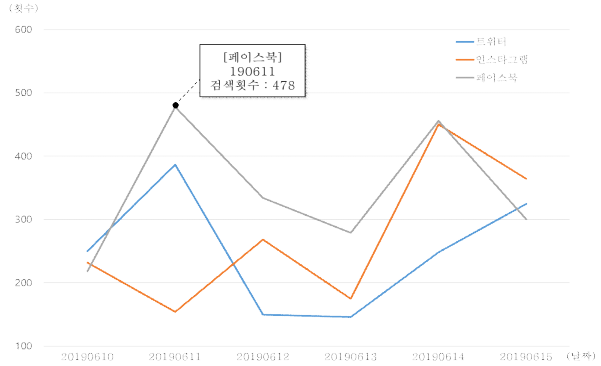
(그림 3)은 SNS API를 이용하여 크롤링을 진행하는 프로그램의 작동순서를 보여준다. 사용자가 분석을 원하는 키워드를 입력하면 사용자의 접근 권한과 요청정보를 확

3) https://dev.twitter.com/apps

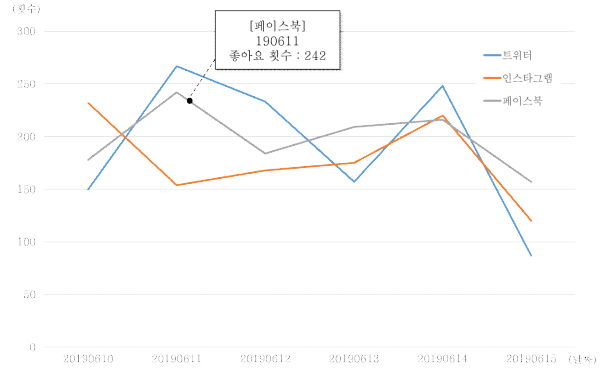
인한다. 액세스 토큰을 설정하고 참조변수를 이용하여 API를 생성함으로 크롤링을 진행할 환경을 설정해준다. 크롤링 환경이 모두 준비되면 원하는 수집 기간을 설정하고 KoNLPy를 이용해 텍스트의 형태소 분석을 진행한다. 본 시스템은 명사단위의 형태소만을 추출하고, 추출된 명사는 DB에 저장한다.

4. 구현 및 결과

웹페이지로 구성된 시스템은 별다른 조작 없이 트렌드 분석을 진행할 검색어를 검색창에 입력하게 된다. 입력된 검색어는 위 3절의 과정을 거쳐 크롤링 된 SNS 글과 비교되어 검색한다. 그 중 의미 있는 키워드만이 추출되며 횟수가 저장된다. 본 연구에서는 명사 단위의 검색어만이 비교 대상이며, 그 외의 형태소는 오류 메시지를 출력하며 초기화면으로 되돌아간다.



(그림 4) 키워드(어벤져스)의 검색 빈도 그래프화



(그림 5) 키워드(어벤져스)의 공감(좋아요) 횟수 표시

저장된 키워드의 횟수는 각 SNS에 따라 서로 다른 색상의 그래프로 표시된다. 그래프의 x축은 일정기간 동안의 날짜를 의미하고 있으며, y축은 키워드의 검색 횟수를 의미하고 있다. 그래프에 마우스 오버 시 정확한 검색 횟수가 나타나게 된다. 기존 연구에서는 하나의 사이트의 검색어 횟수만 비교 가능했지만 본 논문에서는 여러 개의 검색엔진 및 SNS를 하나의 사이트에 포함해 횟수 분석을 진행했다. 이로써 사용자는 각 사이트별 검색 횟수와 트렌드 추이를 한눈에 확인할 수 있다. 검색된 글 중 좋아요 및 공감이 달려있는 글의 개수를 그래프로 표시하여 다른 사용자들의 관심도 또한 확인 가능하다. 검색횟수와 좋아요 횟수, 해시태그 등의 다양한 정보를 보여줌으로써 사용자에게 편리함을 제공한다. 각 그래프의 변화를 보며 누구나 접근 가능한 개방형 SNS(페이스북, 트위터)와 특정 사람들과만 소통이 가능한 폐쇄형 SNS(인스타그램)에 대한 사용자들의 이용 추이를 확인할 수 있다.

5. 결론

본 논문에서는 효과적인 트렌드 분석을 위한 데이터 처리 방법으로 다수의 SNS 사이트를 이용하여 트렌드를 분석하고, 해시태그 및 공감, 공유 등을 이용하는 방법을 제안했다. 제안하는 방식은 여러 SNS의 API를 통해 크롤링 된 글들을 KoNLPy를 이용해 토큰화하여 빈도를 계산하였다. 이 방식으로 얻어낸 결과값은 SNS 종류에 따라 서로 다른 색상의 그래프로 표시되며, 해시태그나 공감(좋아요) 수를 이용하여 트렌드를 확인할 수 있다. 주기적으로 SNS 글을 장기간에 걸쳐 수집하고 본 논문에서 제안한 시스템을 사용하여 용어들의 빈도를 분석한다면, 사용자들이 관심을 갖는 키워드들이 발생해서 확산하고 소멸하는 과정을 살펴볼 수 있을 것으로 기대한다.

감사의 글

이 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음(2019-0-01183)

참고문헌

[1] 김태우, 백동현. "SNS 의 스마트폰 게시글 내용 분석을 통한 사용자의 요구특성 분석." 한국전자거래학회지 17.4 (2012): 197-208.

[2] 윤진영, 김석중, 이범석, 황병연. "트위터에서 추출한 감기 증상의 사회적 신호와 영향요인과의 상관분석." 멀티미디어학회논문지 16.6 (2013): 667-677.

[3] 김민교, 이상준. "개인형 추천 서비스를 위한 사용자 성향분석 엔진." 한국지능정보시스템학회 학술대회논문집 (2013): 190-192.

[4] 고재창, 조근태, 조윤호. "키워드 네트워크 분석을 통해 살펴본 기술경영의 최근 연구동향." 지능정보연구 19.2 (2013): 101-123.

[5] 이윤주, 서지훈, 최진탁. "SNS 텍스트 콘텐츠를 활용한 오피니언마이닝 기반의 패션 트랜드 마케팅 예측 분석." 한국정보기술학회논문지 12.12 (2014): 163-170.